Protecting Privacy When Releasing Search Results from Medical Document Data

Xiao-Bai Li
University of Massachusetts Lowell
xiaobai_li@uml.edu

Jialun Qin
University of Massachusetts Lowell
jialun_qin@uml.edu

Abstract

Health information technologies have greatly facilitated sharing of personal health data for secondary use, which is critical to medical and health research. However, there is a growing concern about privacy due to data sharing and publishing. Medical and health data typically contain unstructured text documents, such as clinical narratives, pathology reports, and discharge summaries. This study concerns privacy-preserving extraction, summary, and release of information from medical documents. Existing studies on privacy-preserving data mining and publishing focus mostly on structured data. We propose a novel approach to enable privacy-preserving extract, summarize, query and report patients' demographic, health and medical information from medical documents. The extracted data is represented in a semi-structured, set-valued data format, which can be stored in a health information system for query and analysis. The privacy preserving mechanism is based on the cutting-edge idea of differential privacy, which offers rigorous privacy guarantee.

1. Introduction

Medical documents and other unstructured data, such as clinical narratives and discharge summaries, are essential for documenting interactions between patients and healthcare providers. These clinical and medical texts are typically embedded in an electronic medical records (EMR) system. They contain rich information useful for improving clinical decision support and for medical and healthcare research [13][20]. Traditionally, extraction of information from clinical text into a form suitable for analysis and research is done manually by domain specialists. In recent years, there have been significant developments in using natural language processing (NLP) techniques for information extraction from medical documents [19][23].

In order to make patient data available for research and analysis, it is vital to ensure that patient privacy is appropriately protected. To this end, the Health Insurance Portability and Accountability Act (HIPAA) [5][6] has established a set of privacy rules. The HIPA Safe Harbor rule specifies 18 categories of explicitly or potentially identifying attributes – called Protected Health Information (PHI) – that must be removed or altered before the health data is released to a third party. However, a strict implementation of the Safe Harbor rule may be inadequate for protecting privacy or preserving data utility. Studies have shown that the Safe Harbor rule lacks the flexibility to adequately meet the diverse needs of data users; it can be under-protective in some cases and over-protective in others [18][25]. Recognizing this limitation, HIPAA also provides guidelines that enable a statistical assessment of privacy disclosure risk in order to determine if the data is appropriate for release. This study focuses on this aspect of the HIPAA principle.

Along the line of the statistical approach, there is a large body of research on privacy-preserving data sharing and publishing, most of which focus on structured data [1][11]. Privacy models such as *k*-anonymity [24], *l*-diversity [17], *t*-closeness [14], differential privacy [8][9], and clustering-based anonymization approaches [15][16] have been proposed to formalize privacy protection requirements. Various methods and algorithms have been developed to anonymize structured data to satisfy the requirements in the aforementioned privacy models [1][11].

In spite of this richness of research in data privacy, its application in medical domains lags behind in some aspects. Medical data typically contain text documents. In such cases, identity information is embedded in the textual contents, where anonymization techniques designed for structured data are not readily applicable. Thus, the majority of privacy research in sharing and releasing information in medical documents has followed the Safe Harbor rule directly, focusing on the automatic detection of PHI attributes in the documents [18][21][25]. The identified PHI values are then simply removed from or encrypted in the released text. Studies have shown that such simple de-identification strategies lack the flexibility to adequately meet the

URI: http://hdl.handle.net/10125/50363
ISBN: 978-0-9981331-1-9

diverse needs of data users; they can be under-protective (i.e., not satisfying privacy requirements) in some cases and over-protective (i.e., resulting in poor data utility) in others [18][25]. There is a lack of research on how to provide adequate information for identified PHI in a privacy-preserving manner (other than simple removal) and how to cope with non-PHI but potentially identifying information to improve privacy protection and data utility.

In this paper, we study privacy issues related to releasing summary and query information from patient medical documents. We propose a novel approach to extract and release patients' demographic, health and medical data from clinical text. The extracted data is represented in a semi-structured, set-valued data format, which is then used for privacy-preserving query and analysis. Our privacy mechanism is designed based on the differential privacy framework.

The main contributions of this research are: (1) We examine a problem that has not been formally studied in the literature, which is releasing PHI related summary information and search query results from medical documents. Existing privacy-preserving techniques for releasing structured data or PHI-removed data are not readily applicable to the problem. (2) We propose a novel approach to release patients' demographic and health information from medical documents. The privacy preserving mechanism is based on the leading-edge idea of differential privacy, which offers rigorous privacy guarantee. (3) We conduct an experimental study that demonstrates the effectiveness of the proposed approach.

This paper is organized as follows. We review related work in data privacy and health informatics in Section 2. We then demonstrate in Section 3 the privacy and data quality problem with the current Safe Harbor practice, using an illustrative example. In Section 4, we present the proposed approach for summarizing medical documents and releasing search query results with privacy guarantee. In Section 5, the results of an experimental study are provided. We conclude our paper and provide future research directions in Section 6.

2. Related Work

In analyzing privacy disclosure risk, the literature typically recognizes two types of disclosure [7]: (a) *identity disclosure* (or *re-identification*), which occurs when an adversary is able to match a record in a de-identified dataset to an actual individual; and (b) *attribute disclosure*, which occurs when an adversary is able to predict the sensitive value(s) of an individual record, with or without knowing the identity of the individual. Related to the two types of disclosures, the attributes of data on individuals can be classified into three types. We discuss them in the context of medical and health data, with respect to the HIPAA-defined PHI categories, as listed in Figure 1 (from pp. 82818-82819 in [5]).

1. Names
2. *Locations*: All geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial 3 digits of a zip code if the correspond area contains more than 20,000 people.
3. *Dates*: (i) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death. (ii) All ages over 89 and all elements of dates (including year) indicating such an age.
4. Telephone numbers
5. Fax numbers
6. E-mail addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web Universal Resource Locators (URLs)
15. Internet Protocol (IP) address numbers
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

Figure 1. Protected Health Information (PHI) Defined by HIPAA

The first type is *explicit identifier* (EID), which are PHI attributes that can be used to directly identify an individual, such as name, phone number and email address. It is clear from Figure 1 that all PHI categories are EIDs except category 2 (locations) and category 3 (dates). HIPAA requires that EIDs be removed or encrypted in the released data. The second type is *quasi-identifier* (QID), which do not explicitly reveal identities but may be linked to external data sources to eventually identify an individual. QIDs include some PHI attributes such as date of birth, admission date, and zip code; they also include some non-PHI

attributes such as age, gender and race. Sweeney [24] found out that 87% of the population in the US can be uniquely identified with three QID attributes – gender, date of birth, and 5-digit zip code – which are accessible from voter registration records available to the public. The third type is *sensitive attributes* (SAs), which contain private information that an individual typically does not want disclosed, such as sexual orientation and personal financial information. In the context of health and medical data, SAs are health and medical information (HMI) such as symptoms, test results, diagnoses, diseases, medications and procedures. None of the items listed in Figure 1 is HMI. That is, HIPAA does not provide guidelines on how to protect SA/HMI information; instead, the basic idea of HIPAA Safe Harbor rule is to protect privacy by preventing identity disclosure.

Most of data privacy studies assume that data is stored in well-defined relational databases. A major line of research has focused on devising principles to establish the requirements of privacy protection and to form criteria for assessing privacy risks. A well-known principle is *k*-anonymity [24], which requires that each individual record in a dataset should be indistinguishable from at least $k - 1$ other records with respect to the QID attribute values. The *k*-anonymity approach focuses on re-identification risk only and does not consider attribute-disclosure risk. To address attribute disclosure, a privacy principle called *l*-diversity has been proposed [17]. The *l*-diversity principle requires that an SA attribute should include at least *l* well-represented values in the *k*-anonymized data. Another privacy principle, called *t*-closeness [14], addresses the issue by further considering the overall distribution of the SA values. It requires that, for each group, the distance between the distributions of the SA values in the group and the overall distribution of the SA values cannot be larger than a threshold value *t*. The *l*-diversity and *t*-closeness principles typically assume that there is a single SA attribute or several pre-defined SA attributes, which is not a realistic scenario for text data. Medical text documents typically have a large number of unstructured (not pre-defined) SA/HMI attributes. It is essentially impossible to apply the idea of *l*-diversity or *t*-closeness for medical document data.

The *k*-anonymity, *l*-diversity and *t*-closeness approaches all depend on some assumptions about the adversary's auxiliary information regarding individual targets. When the assumptions do not hold, these approaches may not work well [9]. To overcome this limitation, Dwork [8][9] introduces the notion of differential privacy. Intuitively, differential privacy ensures that the released information about a dataset is essentially the same whether or not an individual's data was included in the dataset. In other words, there is virtually no additional privacy disclosure risk if the individual opts in to the dataset. Differential privacy is defined independent of any auxiliary information assumption. Thus, it provides the most rigorous privacy guarantee among existing approaches. On the other hand, differential privacy requirements often result in significant information loss in the released data, which limits its applicability. A recent survey found that application of differential privacy to the medical and health domain remains an unexplored research area [4].

Unlike privacy research in the structured data, where numerous techniques have been proposed and developed, privacy protection approaches for sharing information in medical documents have mainly based on the Safe Harbor principle, focusing on the detection and removal of PHI items from the documents. Meystre et al. [18] and Uzuner et al. [25] have reviewed more than a dozen state-of-the-art techniques in the field, all of which follow the Safe Harbor approach and none takes statistical approach that is common in the privacy research for structured data.

Thus, there is a lack of interaction between the study in de-identification for medical documents and that in anonymization for structured data. To integrate these two research streams, existing de-identification techniques need to be extended beyond the HIPAA-defined PHI fields. On the other hand, anonymization techniques designed for structured data, such as differential privacy, need to be adapted to take advantage of the rich sematic information embedded in the textual contents.

3. Privacy and Data Utility Problem with Medical Documents

Given a collection of patient medical documents, our first task is to extract relevant data elements and assign them to the three categories described earlier: EID, QID and HMI. This task involves information extraction and classification. There exist many NLP techniques in medical and health informatics to perform this task [19]. We adopt some existing techniques, to be described later, for this task. After information extraction and classification, EID data will be removed or encrypted, following the HIPAA rule. The QID and HMI data will be stored in a semi-structured scheme for query and analysis. In this scheme, QID will be stored in a standard table, one record for each document. HMI will be stored in a set-valued format, where each set of terms and values that appear in a document is listed together and associated with the QID values of the same document. Such a

scheme is supported by many health information systems that enable the use of EMR data for decision support and health and medical research, such as the i2b2 system [20] and the Vanderbilt research data warehouse system [3]. Our work examines privacy protection issues related to the search and release of information from this scheme.

1. Visited on 4/5/2009. Male, 24 year old. Feeling sore throat, fever, headache, fatigue…
2. Mr. Brown's daughter is 9 year old. Visited on 4/13/2009…Having runny nose, sore throat, fever, headache…
3. Admitted on 4-21-2009, patient is a 9 year old female. Having runny nose, sore throat, diarrhea, fever.
4. Amy is 17 year old. Having fever, joint pain, nausea, sore throat…Visited 5/14/2009.
5. Admitted on 6/7/2009, the 88 year old man is complaining chills, body pain, sore throat, fatigue, fever…

Figure 2. An illustrative example of five clinical notes

No.	Visit Month	Visit Year	Age	Gender	Zip Code	HMI
1	April	2009	24	Male	12301	sore throat, fever, headache, fatigue
2	April	2009	9	Female	12301	runny nose, sore throat, fever, headache
3	April	2009	9	Female	12301	runny nose, sore throat, diarrhea, fever
4	May	2009	17	Female	12302	fever, joint pain, nausea, sore throat
5	June	2009	88	Male	12302	chills, body pain, sore throat, fatigue, fever

Figure 3. Information extracted from the example clinical notes

To describe the idea of our approach, consider a set of five patient clinical notes shown in Figure 2, taken from a hypothetical community hospital. Figure 3 illustrates the scheme that contains extracted QID and HMI values. The first five columns follow a relational database table format (where the additional Zip Code data is obtained from the patient registration). The last column contains a set of HMI terms/values. To comply with the Safe Harbor rule, the hospital can only release the data in Visit Year, Age, Gender and the first three digits of Zip Code, as well as the HMI values. However, this Safe-Harbor-based release can be over-protective. For example, because only the visit year can be released, the important "season" information is lost, which could be crucial for detecting an epidemic disease outbreak. For the same reason, releasing the 3-digit zip code (e.g., 123**), instead of the 5-digit zip code, also causes significant information loss. On the other hand, the Safe-Harbor-based release may be inadequate for privacy protection. For example, it may not be difficult to identify the 88-year-old man (No. 5) in the region who has been hospitalized in 2009, using publically available data.

The proposed mechanism releases information as an output response to a search query using HMI terms. We focus on count query using HMI search terms in this early stage of our study. Even for the count query output only, our approach can provide much more useful information than the Safe Harbor rule. For example, using the HMI search terms {sore throat, fever}, the output can show a count for the following conditions:

Visit Month = 'April',
Visit Year = 2009,
Age = 9,
Gender = 'Female',
Zip Code = 12301,
HMI = {sore throat, fever}.

Since there are two matching records in the example set, the perturbed count will be 2 plus a noise (which will be discussed in the next section). Without loss of clarity, we write the above conditions as:

<April, 2009, 9, F, 12301, {sore throat, fever}>

We can also query with slightly different conditions, <April, 2009, 12301, {sore throat, fever}>, which has three matching records; so the perturbed count will be 3+noise. Moreover, we can also get a perturbed count for <April~June, 2009, 1230*, {sore throat, fever}>, which will be 5+noise. These outputs provide useful information about a flu-like disease that may be spreading in the area during the period (assuming many similar records are found in the entire patient database). Note that it is also possible for the proposed mechanism to output the perturbed count for the records with QID values matching those of record #5 (even though the match is unique), but the noise for the

count is likely to be very large compared to the original count.

4. Differentially Private Data Release

Our proposed method for adding noise is based on the notion of differential privacy [8][9], which is defined bellow:

Definition 1. Given any two datasets D_1 and D_2 that differ in only one record, a perturbation mechanism *M* provides *ε-differential privacy* if for any set of possible outputs *S* of *M* (i.e., $S \subseteq Range(M)$),

$$\Pr[M(D_1) \in S] \le e^{\varepsilon} \times \Pr[M(D_2) \in S]. \quad (1)$$

The parameter ε represents disclosure risk, which is usually controlled to be small so that e^{ε} is close to one. As such, differential privacy guarantees, in a probabilistic sense, that the outputs will be essentially the same with or without any specific individual's participation. This property has a very appealing implication. For example, if the dataset were to be used by a healthcare provider to analyze the demographics of its patient population, then the presence or absence of a patient's record in the dataset will not significantly change the results of the analysis. In this sense, the participating patient's demographic information is well hidden.

For a frequency query (e.g., query for count or histogram), there is a straightforward way to construct a perturbation mechanism that satisfies ε-differential privacy. The mechanism is based on the notion of sensitivity defined below [8]:

Definition 2. For a function *f* over dataset *D* with numeric output, the *sensitivity* of *f* is

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1 \quad (2)$$

for all D_1, D_2 differing in at most one record.

In other words, the sensitivity is the maximum change in the value of *f* when any single record of *D* changes. To add noise for a numeric output, it is convenient to use a Laplace distribution. The Laplace distribution with a scale parameter σ, $Laplace(\sigma)$, has a density function of $p(x \mid \sigma) = (1/2\sigma)e^{-|x|/\sigma}$. With this distribution, we have the following result [9]:

Result 1. *For a numeric function f, a perturbation mechanism that adds noise with a* $Laplace(\Delta f / \varepsilon)$ *distribution to the output satisfies ε-differential privacy.*

To be rigorous, for an integer-valued output, a geometric distribution (instead of the Laplace) should be used for perturbation [12], but this subtle difference is not considered important in the literature. When *f* represents a count query, sensitivity $\Delta f = 1$ since the count can differ at most by one due to the addition or removal of one record. Therefore, for a count query *f*, the perturbation mechanism

$$M(D) = f(D) + Laplace(1/\varepsilon) \quad (3)$$

provides ε-differential privacy.

Given a set of medical documents, our approach first extracts EID-, QID-, and HMI-related terms. There is no existing information extraction system that can effectively extract all these terms. We have adapted two open-source systems to perform this task: the Stat De-id system [26] and the cTAKES system [23]. Stat De-id treats capturing EID and QID terms as a multi-class classification task and uses a support vector machine (SVM) technique to classify a term as an EID or QID category. The cTAKES system is a natural language processing system specialized in medical text domain. It combines rule-based and machine learning techniques aiming at information extraction from medical documents.

We do not use these two systems directly to extract EID, QID, and HMI terms. Instead, we took the basic classifier components from the two systems to build a set of independent base classifiers (e.g., rule-based classifier, SVM-based classifier, conditional random field (CRF)-based classifier, etc.). These base classifiers classify the terms in medical documents into one of the four categories: EID, QID, HMI, or OTHER. The results of the base classifiers are then fed into an ensemble classifier to produce the final combined result. For example, in the combined result for record 2 in Figure 2, "Mr. Brown" will be classified as an EID. Similarly, "9 year old" and "daughter" (which implies female) will be recognized as QIDs. Words such as "sore throat," "fever," and "fatigue," will be classified as HMI.

When there are conflicts between base classifiers, the ensemble classifier resolves the conflicts based on privacy priority. For example, an EID causes direct disclosure of an individual's identity and thus has the highest privacy priority among the four categories. If a term is recognized as an EID by any classifier, it will be classified as an EID and removed from the text, even though it is recognized as a QID or HMI by all the other classifiers. As an example, if a base classifier recognizes "White" as an EID (patient name) and the other base classifiers consider it as a QID (race),

"White" will be classified as an EID and removed from the anonymized text. This enables maximum protection for the EID attributes. Similarly, if a term is classified as a QID by one classifier but as an HMI by the other classifiers, it will be classified as a QID because a QID value is subject to change (e.g., zip code 123**) while an HMI value will remain unchanged in the anonymized text.

After extraction, EID values are removed or encrypted. The QID and HMI data are populated into a scheme exemplified in Figure 3. Each row in the scheme can be viewed as a transaction in the context of association rule mining, and each value or term can be viewed as an item. Therefore, the Apriori algorithm [2] can be applied to find frequent itemsets. The minimum support count for the frequent itemsets (i.e., the number of transactions containing the itemsets) can be considered as a privacy parameter (in addition to ε). This parameter can be controlled by the data owner but unknown to the data user. It can be set to a relatively small value because the count will be perturbed before it is released.

1. For a set of medical documents, extract EID, QID and HMI terms and values.
2. Remove or encrypt EID values. Load QID and HMI values into a table D where the HMI field allows a set of multiple terms or values.
3. Run the Apriori algorithm on D to find all frequent itemsets that contain at least an HMI value.
4. For a count query $f(D)$ involving a set of HMI value, obtain the count result from the output of Step 3. Perturb the result using Equation (3).

Figure 4. Computational Procedure

The entire computational procedure for our proposed approach is summarized in Figure 4. In terms of computational complexity, Steps 1, 2 and 3 can be preprocessed, so the real time computation for a query is very fast. Also, the Step 3 computation is faster than that of the classical Apriori algorithm because the itemset not containing any HMI term can be removed immediately at each Apriori iteration. So, the computation is efficient even if it is necessary to re-run Step 3 (due to, for example, a change in the support parameter).

5. Experimental Evaluation

To evaluate the proposed approach and compare it with the Safe Harbor approach, we have conducted an experimental study using real patient document data.

The Informatics for Integrated Biology and the Bedside (i2b2) project has obtained multiple sets of medical documents from healthcare organizations and made them available for research (https://www.i2b2.org/NLP/DataSets). We used four of the datasets for the experiment, all of which are medical discharge summaries. The first set is related to a clinical-text de-identification challenge competition. The second set was initially used for evaluating document classification techniques. The third set was used for extraction of medication information from clinical text. The fourth set was used for a challenge competition to extract medical concepts, assertions and relations. Because all of the datasets are medical discharge summaries, the elements of information contained in different datasets are similar, most including patient name, admission and discharge date, age, gender, hospital, symptoms, test result, diagnoses, diseases, medications, and so on. Thus, we merged the four sets into a single set, resulting in 2,867 text records. After extracting the QID and HMI values from the text, we found that there were very few zip code and/or location values in the data that we could use for the experiment. Therefore, we focused on the query results involving the visit year and month data, which appear in nearly all records. Recall that the Safe Harbor rule prohibits releasing patient visit month data.

The privacy protection level is naturally measured by the parameter ε. Clearly, the smaller the ε value, the better the privacy protection the mechanism offers. In terms of data utility, since the count query result can be regarded as an itemset count, we use an itemset-related measure in the literature [10][22], called *relative error*, which is defined as

$$\text{Relative Error} = \frac{1}{|I|}\sum_{i\in I}\frac{|\tilde{n}_i - n_i|}{n_i}, \qquad (4)$$

where I represents the set of all frequent itemsets with support count larger than the specified threshold value; n_i and $\tilde{n}_i$ are respectively the original and perturbed count of the ith frequent itemset. Since an itemset must contain at least an HMI item, the relative error measures the error rate for the results of the queries having at least an HMI term while satisfying the minimum support count requirement.

We set parameter ε to five different values: 0.1, 0.2, 0.3, 0.4, and 0.5, which are in general more conservative (i.e., with stronger privacy protections) than commonly used ε values in differential privacy research. To evaluate the performance at different frequency levels, we set minimum support count to five values: 10, 20, 30, 40, and 50. The results of the perturbation algorithm vary slightly with different random number seeds. Therefore, for each scenario the

algorithm was run five times, each run using a different seed. The average results are reported.

The results of the experiment are shown in Figure 5. It is observed that the error rate decreases as the privacy risk (ε) increases, which is expected. Furthermore, the error rate decreases as the support count increases. This also makes sense because the frequency count for the selected itemsets (the denominator in Equation 4) becomes larger when the support is increased. The added Laplace noise, however, is independent of the support. When ε is small, its value is approximately the odds that the output results will be different due to the addition or removal of any record (e.g., when $\varepsilon = 0.1$, the odds is about $e^{0.1} - 1 = 0.105$). Note that the results are based on the queries that allow releasing visit month data, which is prohibited in the HIPAA Safe Harbor rule. Therefore, the proposed approach provides an additional option to Safe Harbor for data release, based on well-grounded assessment of disclosure risk. If the data include other HIPAA-restricted QIDs such as zip code, location, and date of birth, similar analyses can be performed based on our approach.

Privacy Parameter	Support Count				
	10	20	30	40	50
$\varepsilon = 0.5$	0.149	0.078	0.056	0.042	0.034
$\varepsilon = 0.4$	0.184	0.096	0.070	0.055	0.046
$\varepsilon = 0.3$	0.246	0.132	0.095	0.074	0.060
$\varepsilon = 0.2$	0.350	0.187	0.131	0.104	0.086
$\varepsilon = 0.1$	0.731	0.394	0.288	0.223	0.177
Month Estimated	0.510	0.433	0.436	0.461	0.436

Figure 5. Results of Relative Error

Because the problem we study is new to the literature, there are no existing techniques that can be compared directly. We have assumed a scenario where the released output is Safe Harbor compliant (i.e., without month), but the data user attempts to estimate the month value for the query output with a probability proportional to the marginal distribution of the month values. The month values in the dataset are distributed unevenly, ranging from 4% for the least frequent month to 13% for the most frequent month. The error results under this Safe Harbor scenario are shown on the last row of Figure 5 (labeled "Month Estimated"). It is observed that the resulting error rates are in general much higher than those from our approach, particularly when the count becomes large. Therefore, if the month information is important, it is worthwhile to consider using the proposed approach. If the dataset contains location data such as zip code, the proposed approach can also be applied similarly to obtain count results for location values at a more detailed level than that allowed by Safe Harbor (e.g., 5-digit zip code rather than the first 3 digits only).

Clearly, the proposed approach outperforms the HIPA's Safe Harbor rule in this experimental study. Safe Harbor applies the same standard for de-identifying data, which expectedly causes under-protection for some data but over-protection for others because disclosure risks in different data are different. Recognizing this limitation, HIPAA also provides guidelines that enable statistical assessment and control of privacy disclosure risk. Our work follows this line of approach in HIPAA. The proposed approach integrates medical informatics techniques with differential privacy, a statistical perturbation method. This allows the user to have more flexibility in dealing with different disclosure scenarios.

6. Conclusion and Future Directions

In this study, we investigate the privacy issues related to summary and release of information from patient medical documents. We propose a novel approach to extract and release patients' demographic, health and medical data from medical documents. Our approach is based on the well-grounded notion of differential privacy, which offers rigorous privacy guarantee. Our experiments show that the proposed method outperforms the HIPAA Safe Harbor approach. Therefore, the proposed approach may be a promising alternative to Safe Harbor for releasing information from medical document with appropriate privacy protection.

Due to increasing applications of medical data sharing in practice, there is a rising concern that patient privacy is being compromised. The proposed approach will reduce the disclosure risks of individuals from anonymized data, while improving the utility of the data. This should alleviate patients' concerns about loss of privacy and confidentiality and increase their willingness to participate in research that uses patient data. The proposed approach will also reduce organizations' concerns about potential privacy violations and enable organizations to safely share and publish high-quality data for legitimate research and analysis.

One limitation of this study is that the proposed approach was tested on only a relatively small dataset for proof of concept. This dataset might not be an ideal representation of various patient populations. It will be more helpful if some larger datasets are used for experimental evaluation. In order to compare the

proposed approach with the Safe Harbor approach, the PHI values in the original data need to be more detailed than those restricted by Safe Harbor (e.g., date of birth instead of year of birth, and 5-digit zip code instead of 3-digit zip code). Due to data holder's privacy concern, it is very difficult to obtain data with more detailed information than that allowed by Safe Harbor. Future research will obtain more and larger datasets to further validate the proposed approach.

Another limitation is that the proposed perturbation mechanism only applies to query output, not to the original data. It is well-known that output perturbation is vulnerable to the same repeated query attack because in this case the independently added noises will eventually be averaged out, revealing the true value of the query output. Consequently, protection provided by noise perturbation will no longer be effective, causing disclosure of individuals' sensitive information. A simple solution to this problem is to limit the number of repeated query. Various other methods have also been proposed to address this problem, but they all cause considerable deterioration in output quality [9]. For our problem which deals with unstructured data, however, it is possible to add noise in between the text input and the query output. For example, the noise addition may be performed in the information extraction stage. Future research will investigate viable approaches along this direction. It is also possible to release the extracted data directly in a set-valued format using differential privacy, but a fairly large amount of noise may be required for such a direct release. It appears that a relaxed notion of differential privacy may be necessary for this task.

7. References

[1] Aggarwal, C.C., and Yu, P.S. (eds.) *Privacy-Preserving Data Mining: Models and Algorithms*, Springer, New York, 2008.

[2] Agrawal, R., and Srikant, R. "Fast Algorithms for Mining Association Rules in Large Databases," *Proceedings of 20th International Conference on Very Large Databases* (VLDB 1994) Morgan Kaufmann, San Francisco, 1994, pp. 487-499.

[3] Danciu, I., Cowan, J.D., Basford, M., Wang, X., Saip, A., Osgood, S., Shirey-Rice, J., Kirby, J., Harris, P.A. "Secondary Use of Clinical Data: The Vanderbilt Approach," *Journal of Biomedical Informatics* 52, 2014, pp. 28–35.

[4] Dankar, F.K., and El Emam, K. "The Application of Differential Privacy to Health Data," *Proceedings of 5th International Workshop on Privacy and Anonymity in the Information Society*, ACM Press, New York, 2012.

[5] Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information," *Federal Register* 65(250), December 28, 2000, pp. 82462-82829.

[6] Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information," *Federal Register* 67(157), August 14, 2002, pp. 53181-53273.

[7] Duncan, G.T., and Lambert, D. "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics* **7(**2), 1989, pp. 201-217.

[8] Dwork, C. "Differential Privacy," *Proceedings of 33rd International Colloquium on Automata, Languages and Programming, Part II* (ICALP 2006), Springer, Berlin, 2006, pp. 1–12.

[9] Dwork, C. "A Firm Foundation for Private Data Analysis," *Communications of the ACM* 54(1), 2011, pp. 86-95.

[10] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. "Privacy Preserving Mining of Association Rules," *Information Systems* 29(4), 2004, pp. 343-364.

[11] Fung, B.C.M., Wang, K., Chen, R., Yu, P.S. Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* 42(4), 2010, pp.14:1-53.

[12] Ghosh, A., Roughgarden, T., and Sundararajan, M. "Universally Utility-Maximizing Privacy Mechanisms," *Proceedings of 41st Annual Symposium on Theory of Computing* (STOC 2009), ACM Press, New York, 2009, pp. 351-359.

[13] Jensen, P.B., Jensen, L.J., Brunak, S. "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," *Nature Reviews Genetics* 13(June), 2012, pp. 395-405.

[14] Li, N., Li, T., and Venkatasubramanian, S. "*t*-Closeness: Privacy Beyond k-Anonymity and *l*-Diversity," *Proceedings of 23rd IEEE International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, 2007. pp. 106-115.

[15] Li, X.-B., and Sarkar, S. "Class-Restricted Clustering and Microperturbation for Data Privacy," *Management Science* 59(4), 2013, pp. 796–812.

[16] Li, X.-B., and Sarkar. S. "Digression and Value Concatenation to Enable Privacy-Preserving Regression," *MIS Quarterly* 38(3), 2014, pp. 679-698.

[17] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. "*l*-Diversity: Privacy Beyond *k*-Anonymity," *Proceedings of 22nd IEEE International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, 2006, pp. 24-35.

[18] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., and Samore, M.H. “Automatic De-identification of Textual Documents in the Electronic Health Record: A Review of Recent Research,” *BMC Medical Research Methodology* 10, 2010, Article 70, 16 pages.

[19] Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F. “Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research,” *IMIA Yearbook of Medical Informatics*, 2008, pp. 128-144.

[20] Murphy, S.N., Weber, G., Mendis, M., Chueh, H.C., Churchill, S., Glaser, J.P., Kohane, I.S. “Serving the Enterprise and beyond with Informatics for Integrating Biology and the Bedside (i2b2),” *Journal of American Medical Informatics Association* 17(2), 2010, pp. 124-130.

[21] Murphy, S.N., Gainer, V., Mendis, M., Churchill, S., Kohane, I. “Strategies for Maintaining Patient Privacy in i2b2,” *Journal of the American Medical Informatics Association* 18(Suppl 1), 2011, pp. :i103–i108.

[22] Rizvi, S.J., and Haritsa, J.R. “Maintaining Data Privacy in Association Rule Mining,” *Proceedings of the 28th Very Large Data Base Conference*, VLDB Endowment, IBM Almaden Research Center, San Jose, CA, 2002, pp. 682-693.

[23] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., and Chute, C.G. “Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications,” *Journal of the American Medical Informatics Association* 17(5), 2010, pp. 507–513.

[24] Sweeney, L. “*k*-Anonymity: A Model for Protecting Privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5), 2002, pp. 557-570.

[25] Uzuner, O., Luo, Y., and Szolovits, P. “Evaluating the State-of-the-Art in Automatic De-identification,” *Journal of American Medical Informatics Association* 14(5), 2007, pp. 550-563.

[26] Uzuner, O., Sibanda, T., Luo, Y., and Szolovits, P. “A De-identifier for Medical Discharge Summaries,” *Artificial Intelligence in Medicine* 42(1), 2008, pp. 13-35.